Admixture Correction in the Outgroup f3 Statistic

Presented by Nita Tunga

In partial fulfilment of the requirements for graduation with the Dean's Scholars Honors Degree in the Department of Mathematics
University of Texas at Austin

May, 2018

Prof. Jennifer Mann Austin, Ph.D
Supervising Professor

Prof. Kathryn Dabbs, Ph.D
Second Reader

Table of Contents

Introduction

Genetic inheritance can be studied within a purely genetic scope. However, this eliminates part of the picture. The field of genetics is often thought of as a natural science with little in common with fields of social science. However, in human genetics and the genetics of the organisms which humans impact, the role of cultural and societal forces cannot be ignored. For instance, lactase is an enzyme used to digest lactose in milk. As such, it is an enzyme whose activity reduces significantly after weaning. Nonetheless, as humans have begun to ingest more dairy products into adulthood, lactase persistence has evolved to enable humans to digest these dairy products.

My research involves mathematically representing the genetic similarity of two populations accurately via the f3 statistic. The outgroup-f3 statistic is useful in understanding a population's genetic history and how genetically related two populations are. It shows how close two populations are compared to a third population that is equally distant genetically from the first two. However, if two populations share a recent genetic interaction with another population, the outgroup-f3 statistic could show those two populations as being closer together than they truly are. This genetic interaction of two or more previously isolated populations interbreeding is referred to as admixture. Admixture skews, or even inhibits, an understanding of those populations' genetic histories.

To avoid this problem, I have attempted to devise a modified version of the outgroup-f3 statistic to ensure an accurate representation of genetic relatedness. For my project, artificial admixture was introduced in six unadmixed human populations. Depending on the relationship between increased contamination and the f3 statistic, we proposed and adjusted solutions for a corrected f3 accordingly.

I tested my proposed corrections by applying it to populations that contain individuals with and without recent histories of genetic admixture. After correcting for the proportion of admixture in the population, I compared this corrected outgroup-f3 statistic to the outgroup-f3 value calculated for the original unadmixed population. The goal of this work is to have a corrected statistic that one can apply to two populations, independent of admixture proportions. Ultimately, this will help us to better understand the evolutionary histories of populations. Moreover, a corrected statistic will aid other researchers as they analyse demographic histories further in the past.

Background

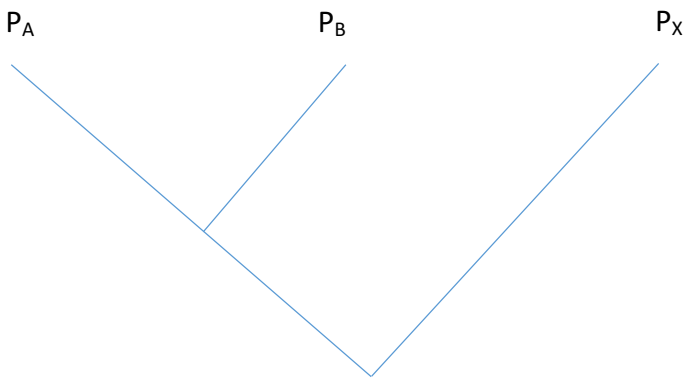
F statistics were first proposed in the paper, “Reconstructing Indian population history”, published in Nature in 2009. In this paper, Reich and colleagues outline the way f2, f3, and f4 statistics can be used to measure genetic drift between two, three, and four taxa respectively. The f3 statistic proposed in this paper is useful for detecting admixture between groups. To summarise, the f3 statistic assumes a null hypothesis of no admixture, which implies a nonnegative f3 statistic. F3 is best used to detect admixture when the time between original split and secondary contact is large, coalescence before admixture is unlikely, and the admixture proportion is close to 50%.

In regard to f3 statistics in particular, Reich et al. propose an equation to be used to measure the genetic drift between three populations, Populations X, A, and B. This equation is, in a simplified form, $f3 = (x - a) * (x - b)$, where *x, a,* and *b* represent allele frequencies in their respective populations. By simplifying the equation, we see that there is a proportional relationship between the f3 statistic and the genetic drift between Populations A and X and Populations B and X. Genetic drift is defined to be the change in allele frequency along a graph edge on a phylogenetic tree. Phylogenetic trees are graphical representations of the genetic relationship between a group of individuals or populations based on physical or genetic characteristics. The length of the branches on the tree often represent the genetic distance, number of genetic differences, between individuals or populations.

More specifically, the calculated f3 statistic is the product of the frequency difference between those populations. This test is useful to see if certain groups have inherited genes from different ancestries. When there has been no admixture, the f3 statistic is expected to be positive. When there has been admixture, the f3 statistic could be

negative. Furthermore, lower f3 values are indicative of less closely related populations, whereas higher f3 values are indicative of more closely related populations. The farther apart two populations are, the smaller the two terms (*x*-*a* and *x*-*b*) in the equation, and therefore, the lower the f3 statistic. Similarly, when two populations are closer together, the two terms in the equation are larger, resulting in a larger f3 statistic.

To better understand what the f3 statistic can be used for, we refer to the figure below.



Here we see that there are two populations that are closer together (Populations A and B), than they are to the third population (Population X). In the context of the equation, $f3 = (x - a) * (x - b)$, we see that we are comparing the allele frequencies in Populations A and B, in relation to the allele frequencies in Population X. If we see how far Population A's allele frequencies are from Population X's allele frequencies and compare this to the distance between Population B's allele frequencies and Population X's, we can evaluate the genetic distance between Population A and Population B. To think about this in a different way, by subtracting out Population A's allele frequencies from those of Population X, we are seeing how much longer or shorter one branch length is compared to the other. Doing so enables us to analyse the distance of each of the three populations in relation to the vertex that connects all three of them. However, if we have an unknown Population Y that

integrates its DNA into both Population A and Population B, it would appear that these two populations are closer genetically than one would expect. In terms of the equation, this would make both terms (x-a) and (x-b) increase or decrease together. As such, the resulting f3 value will be inordinately higher or lower. This is an interesting result if one is concerned with the relationship of Population Y to Populations A and B. However, if you are interested in the genetic relationship of Populations A and B before their admixture with Population Y, this can be a confounding factor.

Nick Patterson was able to work through more of the math behind the F statistics tests, which he documented in his paper "Ancient Admixture in Human History," published in Genetics in 2012. He also discusses the outgroup case, which is further discussed in Maanasa Raghavan's paper, "Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans," published in Nature in 2014. In this paper, the concept of outgroup-f3 statistics is introduced. Outgroup-f3 statistics involve comparing two populations to a third, "outgroup," population, which is equally genetically removed from the other two populations. By doing so, the outgroup population serves as a reference group for measuring genetic relatedness of the populations in question. So instead of looking for admixture between Population X and the other populations, the outgroup-f3 statistic is a measure of the genetic similarity between Populations A and B.

In Benjamin Peter's paper, "Admixture, Population Structure, and F-Statistics," he provides a clear overview of F and D statistics (Genetics, 2017). He also makes the point that f3 statistics can be used as a test for admixture, not just for how closely related two populations are. He also points out that in the history of humans, many of the calculated f3 values are negative, which could show that population phylogenies are not always the best way to discuss human evolution.

F3 statistics have been useful in determining a variety of genetic relatedness questions and are widely used in the field of human population genetics and evolutionary biology more broadly. For instance, outgroup-f3 statistics were used to test relatedness between Levantine and southern Arabian populations to African populations along the Northern and Southern Dispersal Routes out of Africa. Humans evolved in Africa over the past 2 million years. A major dispersal of humans out of Africa occurred around 50 thousand years ago and led to the majority of human genetic variation we see across the world today. Anthropologists and geneticists have long debated whether the primary route out of Africa was the Northern Route or the Southern Route. In "Testing support for the northern and southern dispersal routes out of Africa: an analysis of Levantine and southern Arabian populations," Vyas and colleagues attempted to answer that question using f3 statistics (American Journal of Physical Anthropology, 2017). The Northern Dispersal Route led into Levant, whereas the Southern Dispersal Route led into southern Arabia. By using f3 statistics to see how linked the populations were pairwise, it was found that neither dispersal route was favoured over the other. The Mbuti, a group of people currently living in central Africa, was used as the outgroup population for this test. The results showed that both the Levantine and Arabian populations were equally related to the African population.

The f3 test was taken further and used to show that both the Levantine and Arabian populations shared relatively similar relatedness to non-African populations. Within each region, some groups had more sub-Saharan ancestry, which led to lower f3 values. Another reason for a lower f3 statistic could be an earlier divergence from non-African populations, which would be useful in determining which route was used by earlier populations. The statistic was used to show that both populations were generally equally related to all the

African populations as well. Therefore, the researchers were not able to distinguish which dispersal route was used more.

The f3 statistics have also been used in exploring the relatedness of various subspecies of grapes. In contrast to the previous example of outgroup-f3 statistics, this test used normal f3 statistics to see what sort of admixture has occurred in the history of the grape. While this involves understanding how related two species of grapes are, the primary purpose of this study was to see how to maximally utilise the genetic diversity of grapes. The grape's history of domestication began around 6000-8000 years ago, when the domesticated grape, *Vitis vinifera vinifera*, was cultivated from the wild grape, *Vitis vinifera sylvestris*. The f3 statistic was used to test for mixture between *vinifera* west, *vinifera* east, and *sylvestris* west (f3 = -0.00481); f3 statistics were also used to test for mixture between *sylvestris* west, *vinifera* west, and *sylvestris* east (f3=0.0268).

The researchers found that western *vinifera* is most likely a combination of eastern *vinifera* and western *sylvestris*. Nonetheless, there does not seem to be a genetic transfer between western *vinifera* and western *sylvestris*. This supports that *vinifera* originated in the Near East and underwent introgression into *vinifera* from wild *sylvestris* in Europe.

This analysis found that little of the potential genetic diversity of the grape has been explored. The researchers use this finding to suggest that to overcome the grape's significant pathogen pressures, its genetic diversity must be utilised to its advantage. The domesticated grape contains genetic variation much larger than that of humans, thus making it ideal to manipulate for its polymorphisms and genetic diversity.

Project

The goal of this project is to correct for admixture when calculating the outgroup-f3 statistic so it is an accurate measure of genetic relatedness. I first proposed a similar correction to that used by Lindo et al. for the D statistic.

The D statistic can be used to test for admixture across four populations. In his paper, "Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity," John Lindo proposed a contamination correction to account for similar admixture histories for this statistic (Proceedings of the National Academy of Sciences of the United States of America, 2017). The contamination correction factor Lindo proposes is based on contamination of an ancient genome with modern DNA from a distantly related population, though the one we propose for f3 statistics will be based on the level of artificially induced admixture. Nonetheless, Lindo used a corrected formula to calculate a new D statistic, with admixture corrected for using the contamination correction.

$$D^{*}_{Shuká\ Káa} = \frac{D_{Shuká\ Káa} - cD_{GBR}}{1-c}.$$

$D_{Shuká\ Káa}$ is the contaminated sample's D statistic; D_{GBR} is the D statistic, substituting an individual representative of the population that contaminated Shuká Káa; *c* is the contamination rate. For the f3 statistic, this equation would look like $f3^{*} = \frac{f3 - a * f3_a}{1-a}$, where f3 is the contaminated sample's f3 statistic, $f3_a$ is the f3 statistic with an outgroup as the population that contaminated the original group, and a is the admixture proportion.

Dataset

Our research group utilised population data from North and South American indigenous populations. The first step of this project was to gather usable removed SNPs that were missing in more than 90% of the population, and pruning SNPs based on linkage disequilibrium. I next used the ADMIXTURE program to identify individuals with evidence of European admixture. Populations were then split into three groups: those that had no evidence of European Admixture (Cabecar, Mixe, Surui, Guarani KW, Xaltocan, and Xavante), those where a number of individuals were admixed and a number were not (Jaltocan Hidalgo, Pima, Xaltocan), and those where the entire population had European admixture (Aleut Raff, Algonquin, Cree, Chipewyan, Inupiat, Ojibwa, and Southern US Native American).

Populations into which Admixture was Artificially Introduced	**Population with Admixed and Unadmixed Individuals**	**Admixed Populations on which to Test Correction**
Cabecar	Jaltocan Hidalgo	Aleut Raff
Mixe	Pima	Algonquin
Surui	Xaltocan	Cree
Guarani KW		Chipewyan
Xaltocan		Inupiat
Xavante		Ojibwa
		Southern US Native American

Methods

For my project, I used six completely unadmixed human populations from North and South America - Cabecar, Mixe, Surui, Guarani KW, Xaltocan, and Xavante. I introduced artificial admixture in constant 5% intervals from 5% to 95% admixture from a European population. This was done via a program in R that arbitrary replaced 5 to 95% of the population's genome with the corresponding segment of a European genome. Below is an example of the code used to induce admixture in the population Cabecar using a for-loop.

- ADM=(0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95)
- for j in "${ADM[@]}"; do Rscript admixer.R --file ./final_dataset_cleanest2.vcf --donor Spanish --recip Cabecar --p $j --subs 5 --out final_admix_Cabecar_$j.vcf; done

After simulating admixture in these populations, I obtained outgroup-f3 values for each of these populations and each of the admixture levels within them using the program *popstats*. I also obtained an f3 statistic by swapping out the English population for the Yoruba population, a west African group assumed to be equally distantly related to all these populations, as the outgroup. This outgroup serves as a reference group to compare the desired population and the ingroup to. Karitiana was used as the ingroup for both tests. Then, we can see how increased admixture affects the statistic. This was done using the commands below, where j spanned the admixture proportions mentioned previously:

- python ~/Desktop/project/bin/popstats/popstats.py --file final_admix_Cabecar_$j --f3 --pops C,Karitiana,Yoruba --informative > final_admix_Cabecar_$j_f3.txt

- python ~/Desktop/project/bin/popstats/popstats.py --file final_admix_Cabecar_$j --f3 --pops C,Karitiana,English --informative > final_admix_Cabecar_$j_f3a.txt

Comparing these values to the admixture levels, I was able to re-evaluate the suggested solution as needed. Then by getting an f3 statistic for these populations and setting the outgroup as the population assumed to have contaminated them (English population), I calculated a new f3 statistic, which was hopefully corrected for admixture.

To further test if this correction worked, I took populations that contained individuals with and without admixtured genomes. By correcting for the portion of the population that was admixed, I saw if this corrected f3 statistic matched the unadmixed portion's f3 statistic. I did this in individuals from the Jaltocan Hidalgo, Pima and Xaltocan populations. I then computed a baseline f3 statistic comparing the whole populations, with Karitiana as the ingroup, and Yoruba as the outgroup. After doing so, I got an f3 statistic from the admixed individuals in these populations in relation to Yoruba, and then got an f3 statistic from the admixed individuals in these populations in relation to an English population.

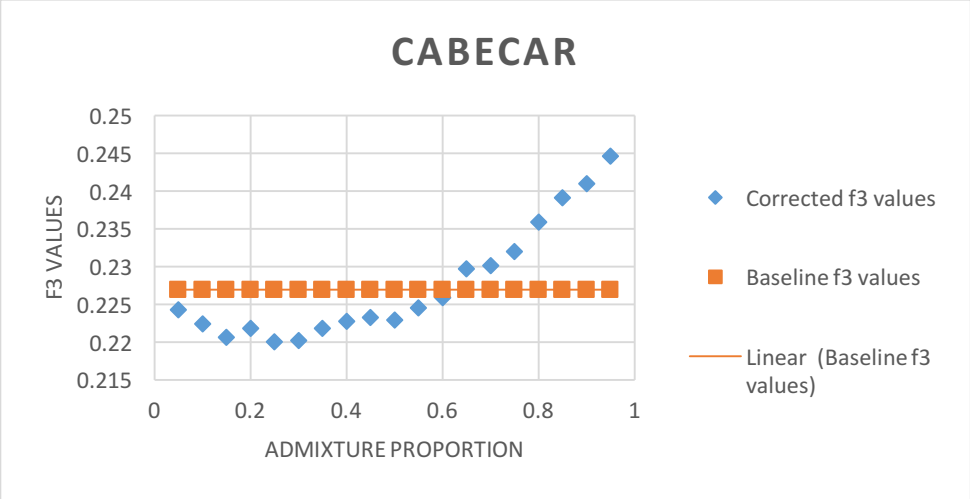
If the f3 statistic was successfully corrected, we could make inferences about the genetic histories of other contaminated populations. I then applied the f3 statistic to the populations Aleut Raff, Algonquin, Cree, Chipewyan, Inupiat, Ojibwa, and Southern US Native American. I obtained the admixture proportion from the amount of European DNA in these individuals. Then I corrected for the f3 statistic by getting an f3 using Yoruba first, and then using English ancestry to compare their genomes to.

Based on preliminary results, the solution could take the form of a corrected equation for outgroup-f3 statistics. On the other hand, it could start with an equation to get

a corrected f3 value, which is then manipulated further. This is where I could come up with a table of values that correspond to different levels of admixture. These “differences” between the semi-corrected f3 and the baseline f3 are then to be subtracted from the semi-corrected f3. Since others attempting to use this correction will not have a baseline f3 for comparison, our goal is to come up with a universal set of differences that can be used depending solely on the admixture levels.

Correction Attempt 1

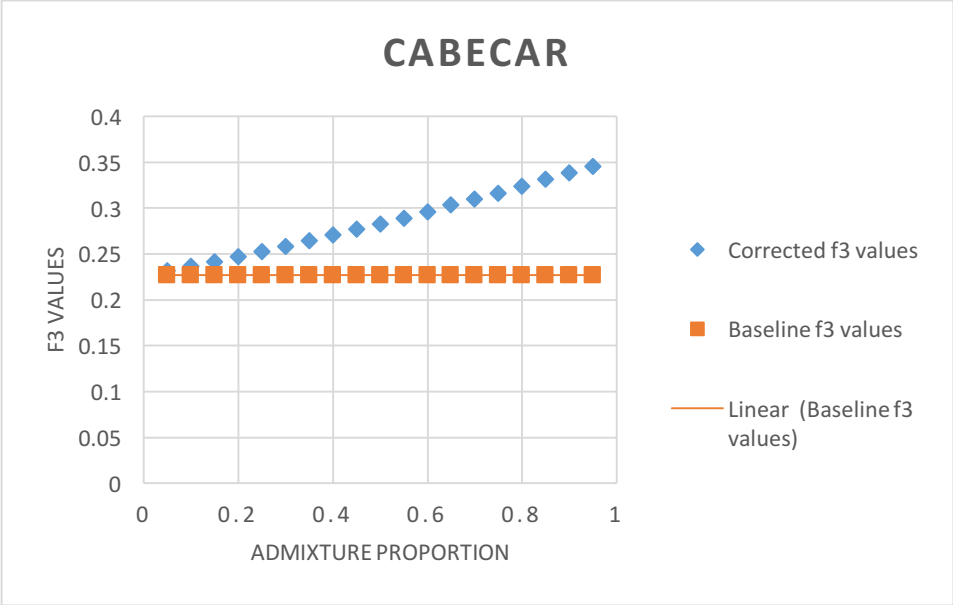
Using the two f3 statistics, I posited a correction equation to get the corrected f3 values to look similar to the baseline f3 values when graphed. A similar correction as that proposed for the D statistic by Lindo was attempted first. However, this was unsuccessful. A new equation was then suggested and tested. This equation took the form of $\big((f3 - f3_a) * a\big) + f3$, where f3 was the statistic calculated with Yoruba as the outgroup, $f3_a$ was the statistic calculated with English as the outgroup, and *a* was the admixture proportion that we introduced into the population. Using these values, I graphed the relation between admixture proportion and the corrected f3 statistic. All the populations' graphs exhibited similar trends. Below is a graph using Cabecar's f3 values to be used as a reference.



Correction Attempt 2

Clearly, the two sets of points are not that similar. As such, I attempted to again correct the equation. Looking at the trend of f3 values dipping around 20-40% admixture levels, it seemed that perhaps I was overcorrecting the f3 values by using f3 values that change with the admixture proportion. As such, I proposed the following equation instead: $\big((f3 - f3_a) * a\big) + f3_{baseline}$, where $f3_{baseline}$ was the value calculated for each of the

populations when there was no artificial admixture introduced. This appeared to at least present a better correlation between admixture and corrected f3 values when graphed. Below is a graph of the newly corrected f3 values plotted against admixture proportions again.

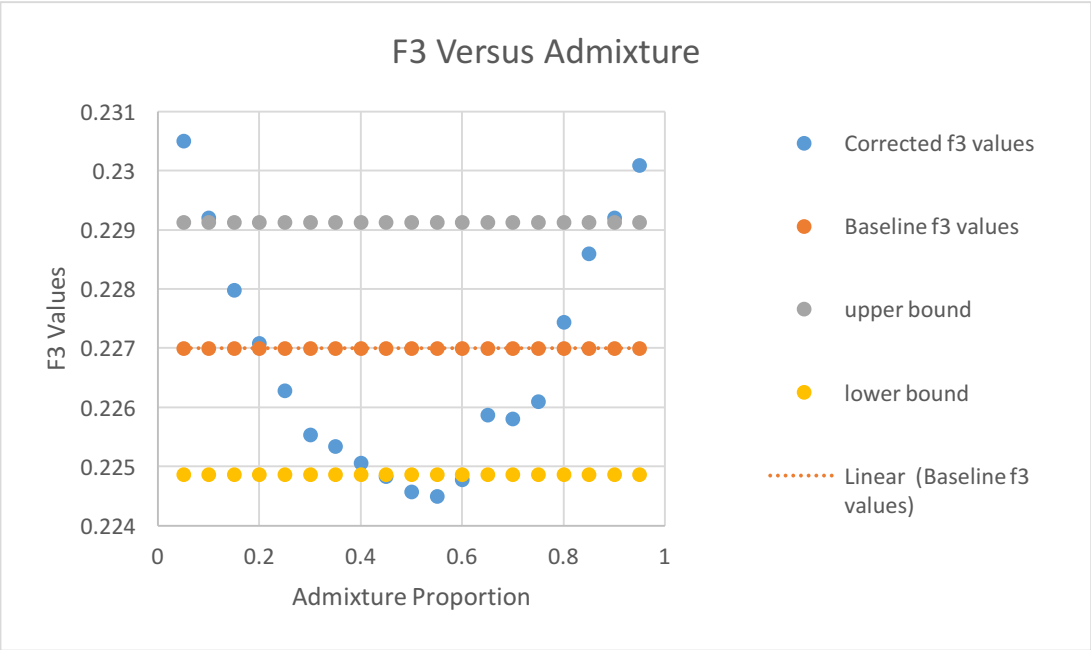


These new f3 values look relatively linear, and as such, I seemed to be on the right track. To further correct the f3 values, I attempted to find the difference between the newly corrected f3 values and the baseline f3 values. I did this for each population, and then found the averages of the differences for each admixture proportion. To the right is a table of the results.

Admixture Proportion	Average Differences
0.05	0.004970747
0.1	0.010104352
0.15	0.015366869
0.2	0.020830249
0.25	0.026399174
0.3	0.032173864
0.35	0.038141322
0.4	0.044222014
0.45	0.050421594
0.5	0.056698452
0.55	0.063145399
0.6	0.06952562
0.65	0.076407052
0.7	0.083150127
0.75	0.089963764
0.8	0.097164007
0.85	0.104699337
0.9	0.111876887
0.95	0.119257046

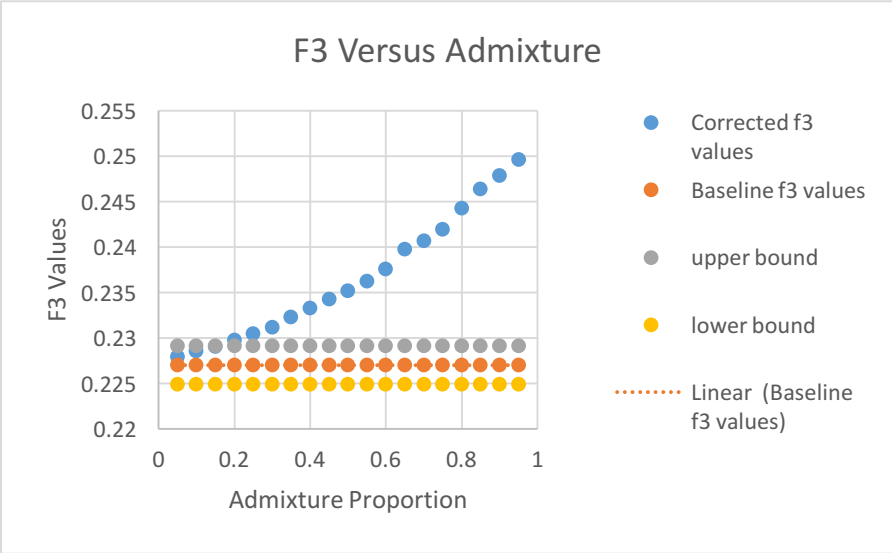
I then plotted the admixture proportions and the average differences, as they looked quite similar. I hoped to see if there was a correlation using a linear relationship. The R^2 value was 0.9973, indicating that there is a significant relationship between these two values. Thus, I attempted to use the equation for the linear regression line as a correction for the f3 values. I used the

values that I had corrected using the equation $\left((f3 - f3_a) * a\right) + f3_{baseline}$ and then subtracted the difference, calculated using the following equation: $y = 0.1272x - 0.005$. Given a certain admixture proportion, I would plug that value in for x in the equation to get the difference to be subtracted from the corrected f3 value. This resulted in a parabolic looking graph of the f3 values plotted against the admixture proportion, shown below (again with the baseline f3 values plotted as a reference for the desired values).



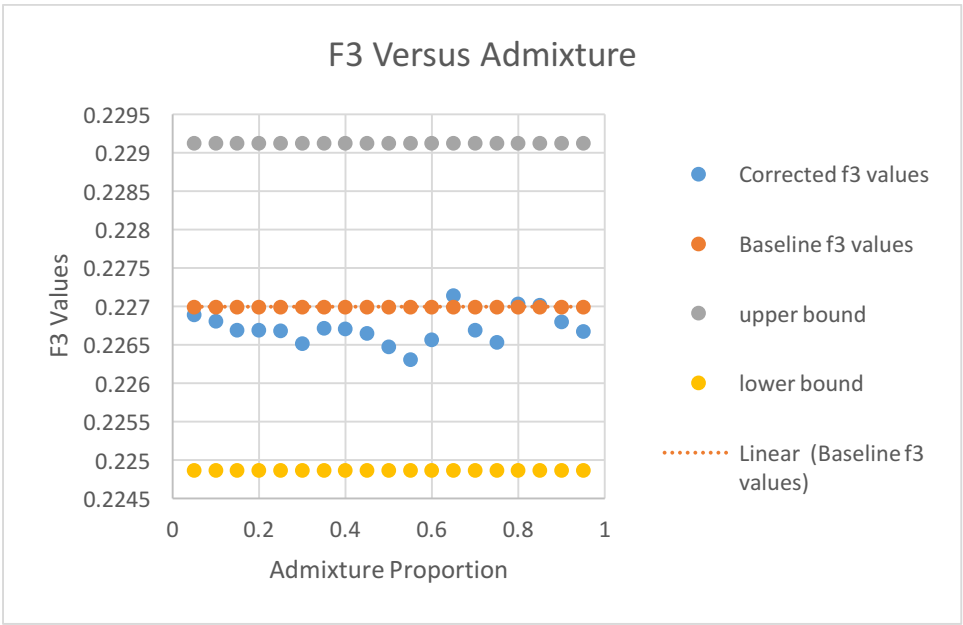
Clearly, this was not an ideal correction of the f3 values again. I attempted to put this in the perspective of the confidence intervals of the baseline f3 values. As such, the upper bound and lower bound representative of one standard deviation above and below the baseline f3 are shown on the graph (the standard deviation was calculated by the *popstats* program used to get the baseline f3 value). Therefore, I attempted to fix the regression equation we had gotten from the average differences. As such, I used the following power equation instead: $y = 0.122x^{1.0859}$. This resulted in an even higher R^2 value of 0.99936, indicating that this equation might work as a correction. Nonetheless, once I used this

equation with the different admixture proportions to subtract from the corrected f3 values, I still had a graph that did not look ideal (below).



Correction Attempt 3:

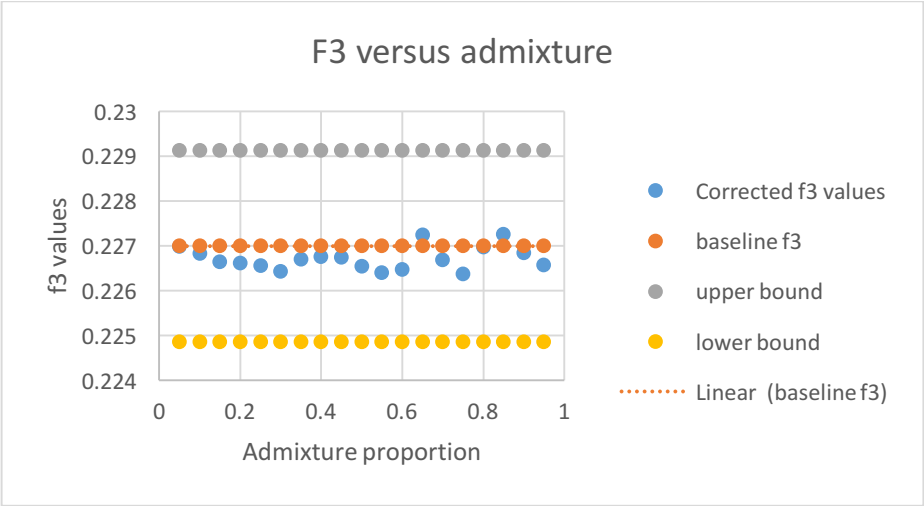
Then, I attempted to just use the average differences to subtract from the corrected f3. I hoped to get these differences for more admixture values where the linear regression line/power line did not match the data well, if this attempt worked. I calculated new f3 values with this correction and got the following graph.



This graph clearly looked a lot better than previous attempts. Furthermore, it was the only solution thus far that yielded corrected f3 values within the bounds of one standard deviation above and below are f3 statistic. Nonetheless, it was not a perfect fit.

To make this graph even better, I got intervals that were closer together (intervals of 1% admixture) between 75 and 85% of admixture**.** This was an area that looked to have a large degree of variance between the baselines and the corrected f3 values. As such, if these new differences that were calculated were better indicators of the difference to subtract from the corrected f3, then I could use these values for the correction.

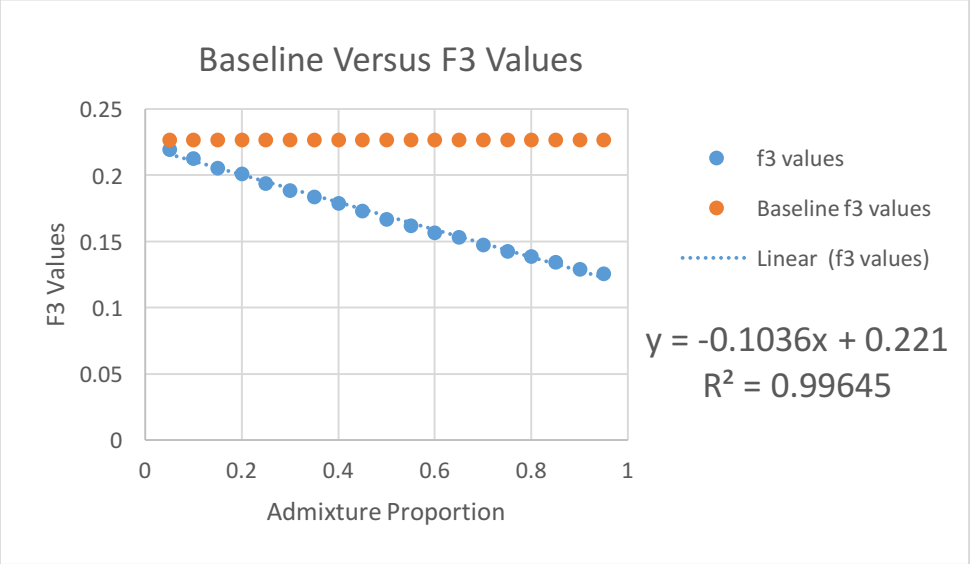
After finding intervals that were closer together, I noticed that this did not significantly impact the correction factor. As such, I tried to use a second order polynomial equation, and got the highest R^2 value yet (R^2=0.99999). Below is the graph when using the quadratic equation to correct the f3 values to baseline.



When I continued with my results, I quickly ran into a snag. I had used the baseline f3 to find a correction equation to get to the baseline f3. In other words, I used the result to force the desired result. However, I was unable to factor out the baseline f3 values to get a correction independent of them.

Correction Attempt 4

As such, I was back to square one and attempted to work with the initial correction equation for f3 ($\left((f3 - f3_a) * a\right) + f3$). I then went back and got the differences between the baseline f3 values and these f3 values. After doing this, I plotted the baseline f3 values against the "corrected" f3 values. There appeared to be a fairly linear trend amongst the f3 values using the equation above, across all six populations. I also noticed that all the f3 values were less than the baseline f3, which reinforced the trend of decreasing f3 values with increased admixture levels. Below is a sample graph from the population Cabecar (with the baseline f3 values in orange, and the preliminarily corrected f3 values in blue). The equation given is for the linear regression line for the preliminarily corrected f3 values.



This led us to believe that we could use the differences between the baseline and the preliminarily corrected f3. After doing this for the six populations, I got the average of the six differences for each admixture level. For instance, I got the average difference for an admixture proportion of 5% across all six populations. After doing so, I used the average differences to get a newly corrected f3 by adding them to the preliminarily corrected f3. I noticed that these new f3 were relatively similar to the baseline f3, though they were not

ideal. As such, I decided it would be beneficial to get a 95% confidence interval for the differences, to see if these confidence intervals of differences would give us something in an appropriate range around the baseline when added to the preliminarily corrected f3.

To do so, I wanted to use a t-test, but the data was not approximately normally distributed. Therefore, I used a Wilcoxon signed-rank test, which is a non-parametrical statistical hypothesis test that allows us to perform a version of the t-test without normally distributed data. It is often referred to as the Wilcoxon T Test. Upon doing so in R, I noticed that the 95% confidence intervals for the differences for each admixture level across the six population would give us a range of differences. Below is a table of these confidence intervals.

Admixture Proportion	**Wilcox Confidence Intervals**
0.05	(0.00750608, 0.02297737)
0.1	(0.01441939, 0.03037686)
0.15	(0.02137041, 0.03316990)
0.2	(0.02569456, 0.04308678)
0.25	(0.03300473, 0.04942423)
0.3	(0.03840717, 0.05224841)
0.35	(0.04300735, 0.05772577)
0.4	(0.04815174, 0.06442764)
0.45	(0.05378068, 0.06854280)
0.5	(0.06017016, 0.07539880)
0.55	(0.06489406, 0.08105843)
0.6	(0.07016284, 0.08537045)
0.65	(0.07382676, 0.09035722)
0.7	(0.07968564, 0.09463120)
0.75	(0.08441375, 0.10056375)
0.8	(0.08823818, 0.10527647)
0.85	(0.09257367, 0.11016625)
0.9	(0.09765934, 0.11534444)
0.95	(0.1012127, 0.1194585)

When the uncorrected f3 values were added to the lower and upper bounds of the confidence intervals, I got intervals for newly corrected f3 values. Once I did this, I noticed

that this interval of f3 values included the baseline f3 values. At first, I hoped to get the baseline f3 values to align with the newly corrected f3 values when using one standard deviation above and below the baseline in conjunction with the confidence interval of newly corrected f3 values. However, the correction using these differences worked well enough that we did not need to consider one standard deviation above and below the baseline f3. Simply using the confidence intervals for the differences to get confidence intervals for corrected f3 values was sufficient as a correction.

I then applied this correction to the natural populations, Jaltocan Hidalgo, Pima, and Xaltocan. I did so by rounding the admixture proportion for these populations to the nearest five hundredths, such that I would be able to use the differences (since we only had these for admixtures that were multiples of 0.05). Upon doing so, I used the confidence intervals for the differences and added the lower and upper bounds to the initial, uncorrected f3 value. Once I did this, I noticed that the baseline f3 statistic fell in this range of new f3 values in the Pima population and in the Xaltocan population. However, this correction did not work for Jaltocan Hidalgo. The range of new f3 values ended up being (0.247343959, 0.262815249), whereas the baseline f3 value was 0.227966338.

Regardless, I then applied this correction to the populations that had admixture, Aleut Raff, Algonquin, Cree, Chipewyan, Inupiat, Ojibwa, and Southern US Native American. I used the confidence intervals for the differences again and rounded the admixture proportion for each population to the nearest five hundredths. Upon doing so, I calculated an interval of f3 values that the baseline f3 is presumed to fall in.

To see if I was able to get a better correction, I plotted the average differences. I was able to use a polynomial regression line since the R^2 values were all above 0.99993. I then got the equation for this curve, which I then used to get a value (using the admixture

proportion as the x value) to add to the preliminarily corrected f3. This results in f3 values that are similar to the f3 values I got from merely adding back in the average difference for the admixture proportion 5% increments. However, they do not fall within one standard deviation of the baseline f3 values, just as adding 5% admixture incremented differences did not yield f3 values that fell within that range either.

As such, I plotted the lower bounds and upper bounds of the 95% Wilcox confidence intervals separately and found regression lines for each. I found that second-order polynomial equations fit the data best (highest R^2 value) and was able to use these equations to add back in the difference to the baseline f3 value. This allowed a continuous correction of the f3 statistic, rather than just at discrete admixture intervals of 5%.

Conclusion

Through the course of this research project, I have developed a crude admixture correction for the outgroup-f3 statistic. By first finding the f3 value of the contaminated population, a "correction factor" can be added back in to bring that value within a ballpark around the baseline f3 statistic. This correction factor comes in the form of a lower bound quadratic equation and an upper bound quadratic equation. When both of these are added to the f3 statistic, the result is a range of f3 values. Comparing these results to the baseline f3 statistics, I conclude that this correction works within a margin of error. Since the correction only worked in two out of the three populations with admixed and unadmixed individuals, we cannot conclude irrefutably that this correction works.

Nonetheless, the correction worked for all admixture levels in all six of the artificially admixed populations ($6 \times 19 = 114$ cases). Therefore, I applied the correction to the seven populations that were completely admixed with European DNA. This resulted in a range of f3 values that resembled appropriate f3 values. However, there is no way to check for which of these seven populations the correction actually worked.

In the future, researchers might be able to fine-tune our correction using data from more populations. For instance, our confidence intervals for the Wilcoxon signed-rank test would likely span a shorter range if there was more data to pull from. Furthermore, it is possible that researchers might be able to further manipulate the postulated equations mentioned previously. Given that Lindo and colleagues were able to find a neat correction equation for the D statistic, it is possible that there exists one for the f3 statistic as well. It was also observed during this project that certain corrections that were suggested worked better at lower admixture proportions. Just as the normal f3 statistic is most accurate under

certain conditions, one of which is that the admixture proportion be close to 50%, it is possible that the outgroup-f3 statistic works best at lower admixture proportions.

Regardless, this correction is useful for researchers hoping to study the genetic relatedness of different populations. In particular, this potential solution is most useful for those hoping to perform outgroup-f3 statistics in populations that have individuals with genetic admixture.

Glossary

- **Admixture**: genetic interaction of two or more previously isolated populations interbreeding
- **D statistic**: a four-population test for admixture
- **F statistic**: measures shared genetic drift between sets of populations
    - Normal f3 statistic: tests for admixture between three populations
    - Outgroup-f3 statistic: proportional to amount of shared genetic history between two populations
- **For-loop**: a control flow statement that specifies iteration to execute a code repeatedly
- **Genetic drift**: the change in allele frequencies in a population over generations as a mechanism of evolution
- **Genetic relatedness**: probability that two individuals share an allele from common ancestry
- **Linkage disequilibrium**: non-random association of alleles at various loci
- **Outgroup**: reference group of organisms not in the populations being studied
- **Phylogenetic trees**: branching diagram representing evolutionary relationships amongst organisms
- **SNPs**: single nucleotide polymorphisms; change in a single nucleotide at a specific genome position

Bibliography

Alexander, David H., et al. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." Genome Research, Cold Spring Harbor Lab, 22 June 2009, genome.cshlp.org/content/early/2009/07/31/gr.094052.109.

Lindo, John, et al. "Ancient Individuals from the North American Northwest Coast Reveal 10,000 Years of Regional Genetic Continuity." PNAS, National Academy of Sciences, 18 Apr. 2017, www.pnas.org/content/114/16/4093.

Myles, Sean, et al. "Genetic Structure and Domestication History of the Grape." PNAS, National Academy of Sciences, 1 Mar. 2011, www.pnas.org/content/108/9/3530.abstract.

Patterson, Nick, et al. "Ancient Admixture in Human History." Genetics, Genetics, 1 Nov. 2012, www.genetics.org/content/192/3/1065.

Peter, Benjamin M. "Admixture, Population Structure, and F-Statistics." Genetics, Genetics, 1 Apr. 2016, www.genetics.org/content/202/4/1485.

Pontussk. "Pontussk/Popstats." GitHub, GitHub, 30 July 2015, github.com/pontussk/popstats.

Raghavan, Maanasa, et al. "Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans." Nature, Macmillan Publishers Limited, 2 Jan. 2014, www.academia.edu/7110954/Upper_Palaeolithic_Siberian_genome_reveals_dual_ancestry_of_Native_Americans.

Reich, David, et al. "Reconstructing Indian Population History." Nature, U.S. National Library of Medicine, 24 Sept. 2009, www.ncbi.nlm.nih.gov/pmc/articles/PMC2842210/.

Vyas, Deven N., et al. "Testing Support for the Northern and Southern Dispersal Routes out of Africa: an Analysis of Levantine and Southern Arabian Populations." American Journal of Physical Anthropology, Wiley-Blackwell, 15 Sept. 2017, onlinelibrary.wiley.com/doi/10.1002/ajpa.23312/full.